

Choice of an Estimate of Genetic Variance from Twin Data

JOE C. CHRISTIAN,¹ KE WON KANG,¹ AND JAMES A. NORTON, JR.²

INTRODUCTION

Twins are frequently used to partition variance of quantitative traits into environmental and genetic components. There is, however, little agreement about how to analyze and present twin data, with some authors constructing heritability estimates and others estimating genetic variance.

The estimate of genetic variance obtained from the twin model is a complex fraction of total population genetic variance: the minimum set of assumptions required for estimating this fraction is explored and the properties of several estimates discussed with the hope that a more rational system of analysis and presentation of twin data may be developed.

The Proposed Model

Table 1 summarizes a general model for estimation of genetic variance using twin data. This model is taken from Haseman and Elston [1], who after Kempthorne [2] used it to estimate total genetic variance from twin data when there are negligible biases in the sampling of twins and negligible effects due to nonrandom mating. Using this model, the four mean squares are independent, and from the expected mean squares we find the following equality:

$$\begin{aligned} E(M_{AMZ} - M_{ADZ}) &= E(M_{WDZ} - M_{WMZ}) = 1/2 \sigma_a^2 + 3/4 \sigma_d^2 \\ &+ (1 - f) \sigma_i^2 + 2 (\sigma_{ge} - \sigma_{ge}^*) + (C_{MZ} - C_{DZ}). \end{aligned} \quad (1)$$

One possible test to determine if this model is valid would be an F' test [3] comparing $(M_{AMZ} - M_{ADZ})$ with $(M_{WDZ} - M_{WMZ})$. However, Cochran [3] advises against using the F' test where a negative sign occurs in the combination of mean squares. We can obtain an equivalent hypothesis by noting that

$$\begin{aligned} E(M_{AMZ} + M_{WMZ}) &= E(M_{ADZ} + M_{WDZ}) \\ &= 2\sigma_a^2 + 2\sigma_d^2 + 2\sigma_i^2 + 2\sigma_e^2 + 4\sigma_{ge}. \end{aligned} \quad (2)$$

Received February 27, 1973; revised July 13, 1973.

This work was supported by the Riley Memorial Association, the John A. Hartford Foundation, NIH contract 71-2307 with the National Heart and Lung Institute, and HL-14159.

¹ Department of Medical Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46202.

² Section of Biostatistics, Department of Psychiatry, Indiana University School of Medicine, Indianapolis, Indiana 46202.

© 1974 by the American Society of Human Genetics. All rights reserved.

TABLE 1

ANALYSIS OF VARIANCE MODEL FOR TWIN STUDIES

Source of Variation	df	Mean Squares	Expected Value of Mean Square	
Monozygotic twins:				
Among pairs ..	$n_{MZ} - 1$	M_{AMZ}	$2\sigma_a^2 + 2\sigma_d^2 + 2\sigma_i^2 + \sigma_e^2 + 4\sigma_{ge} + C_{MZ}$	
Within pairs ..	n_{MZ}	M_{WMZ}	σ_e^2	$- C_{MZ}$
Dizygotic twins:				
Among pairs ..	$n_{DZ} - 1$	M_{ADZ}	$3/2 \sigma_a^2 + 5/4 \sigma_d^2 + (1 + f) \sigma_i^2 + \sigma_e^2 + 2(\sigma_{ge} + \sigma_{ge}^*) + C_{DZ}$	
Within pairs ..	n_{DZ}	M_{WDZ}	$1/2 \sigma_a^2 + 3/4 \sigma_d^2 + (1 - f) \sigma_i^2 + \sigma_e^2 + 2(\sigma_{ge} - \sigma_{ge}^*) - C_{DZ}$	

NOTE.— n_{MZ} = number of monozygotic twin pairs; n_{DZ} = number of dizygotic twin pairs; df = degrees of freedom; σ_a^2 = variance component due to additive genetic effects; σ_d^2 = variance component due to dominant genetic effects; σ_i^2 = variance component due to epistatic genetic effects; σ_e^2 = variance component due to environmental effects; σ_{ge} = covariance between genetic and environmental effects in the same individual; σ_{ge}^* = covariance between genetic effects on one member of a twin pair and environmental effects on the other member of that twin pair; C_{MZ} = covariance among environmental effects between pairs of monozygotic twins; C_{DZ} = covariance among environmental effects between pairs of dizygotic twins; and f = one minus the fraction of epistatic variance manifest within dizygotic twin sets.

An F' test comparing $(M_{AMZ} + M_{WMZ})$ with $(M_{ADZ} + M_{WDZ})$ could therefore be used as a test of appropriateness of the general model. This same comparison was proposed by Kempthorne and Osborne [2] and by Haseman and Elston [1]. Kempthorne and Osborne [2] postulated that a significant value of this ratio would indicate different "competitive forces" for monozygotic and dizygotic (MZ and DZ) twins. These competitive forces could be viewed as environmental variance components unique to each type of twin (σ_{eMZ}^2 = environmental variance component for MZ twins; σ_{eDZ}^2 = environmental variance component for DZ twins). From the nature of these competitive forces it appears that the test should be a two-tailed F' test:

$$F' = (M_{ADZ} + M_{WDZ}) / (M_{AMZ} + M_{WMZ}) \quad (3)$$

or

$$F' = (M_{AMZ} + M_{WMZ}) / (M_{ADZ} + M_{WDZ}),$$

with the larger sum of mean squares as the numerator. The approximate degrees of freedom would be computed as shown in [1] or [3], and the probability would be twice that shown in the usual F tables. If either substantial genetic variance or environmental variance common to both types of twins is present, the power of this test to detect $\sigma_{eMZ}^2 \neq \sigma_{eDZ}^2$ will be low. We therefore recommend performing this test (3) at an increased significance level, perhaps $\alpha = 0.20$.

For data in which there is no evidence for inequality of the total variance of MZ and DZ, at least two assumptions must be made:

$$\sigma_{ge} = \sigma_{ge}^* \quad (4)$$

$$C_{MZ} = C_{DZ}. \quad (5)$$

The first assumption (4) is a simplification of assumption III of Haseman and Elston [1] which was $\sigma_{ge} = \sigma_{ge}^* = 0$, and the second assumption (5) is identical to their assumption IV. In table 2 the twin model is repeated applying

TABLE 2
ANALYSIS OF VARIANCE MODEL FOR TWIN STUDIES,
ASSUMING $\sigma_{ge} = \sigma_{ge}^*$ AND $C_{MZ} = C_{DZ} = C$

Source of Variation	df	Mean Squares	Expected Value of Mean Square	
Monozygotic twins:				
Among pairs ..	$n_{MZ} - 1$	M_{AMZ}	$2 \sigma_a^2 + 2 \sigma_d^2 + 2 \sigma_i^2 + \sigma_e^2 + 4 \sigma_{ge} + C$	
Within pairs ..	n_{MZ}	M_{WMZ}	σ_e^2	$- C$
Dizygotic twins:				
Among pairs ..	$n_{DZ} - 1$	M_{ADZ}	$3/2 \sigma_a^2 + 5/4 \sigma_d^2 + (1 + f) \sigma_i^2 + \sigma_e^2 + 4 \sigma_{ge} + C$	
Within pairs ..	n_{DZ}	M_{WDZ}	$1/2 \sigma_a^2 + 3/4 \sigma_d^2 + (1 - f) \sigma_i^2 + \sigma_e^2$	$- C$

these two assumptions. Equation (1) now simplifies to:

$$E(M_{AMZ} - M_{ADZ}) = E(M_{WDZ} - M_{WMZ}) = 1/2\sigma_a^2 + 3/4\sigma_d^2 + (1-f)\sigma_i^2,$$

the fraction of genetic variance estimated by twin data = G_T , say. We thus have two independent estimates of genetic variance: an among-twin-pair estimate ($\hat{G}_{AT} = M_{AMZ} - M_{ADZ}$) and a within-twin-pair estimate ($\hat{G}_{WT} = M_{WDZ} - M_{WMZ}$).

The hypothesis that twin genetic variance equals zero ($G_T = 0$) may be tested using two separate F ratios:

$$F = M_{WDZ}/M_{WMZ}, \quad (6)$$

$$F = M_{AMZ}/M_{ADZ}. \quad (7)$$

The first of these two ratios (6) would generally be the more powerful because the among-twin-pair mean squares are often much larger than the within-twin-pair mean squares, dwarfing the contribution of genetic variance components

in the second F ratio. By the same reasoning \hat{G}_{WT} would, in most cases, have a smaller variance than \hat{G}_{AT} .

To make use of all of the data available in a single estimate of genetic variance, it would be desirable to combine \hat{G}_{WT} and \hat{G}_{AT} . One possible way of combining \hat{G}_{WT} and \hat{G}_{AT} would be a weighted average based upon the reciprocals of their estimated variances. This estimate (\hat{G}_{MT}) would have minimum variance.

The variances of \hat{G}_{AT} and \hat{G}_{WT} are estimated as follows [4]:

$$\hat{V}(\hat{G}_{AT}) = \hat{V}(M_{AMZ} - M_{ADZ}) = 2 \left[\frac{(M_{AMZ})^2}{n_{MZ} + 1} + \frac{(M_{ADZ})^2}{n_{DZ} + 1} \right], \quad (8)$$

$$\hat{V}(\hat{G}_{WT}) = \hat{V}(M_{WDZ} - M_{WMZ}) = 2 \left[\frac{(M_{WDZ})^2}{n_{DZ} + 2} + \frac{(M_{WMZ})^2}{n_{MZ} + 2} \right], \quad (9)$$

where \hat{V} = estimated variance (for other abbreviations see table 1).

The minimum variance estimate (\hat{G}_{MT}) is then calculated:*

$$\begin{aligned} \hat{G}_{MT} &= \frac{\hat{G}_{AT}/\hat{V}(\hat{G}_{AT}) + \hat{G}_{WT}/\hat{V}(\hat{G}_{WT})}{1/\hat{V}(\hat{G}_{AT}) + 1/\hat{V}(\hat{G}_{WT})} \\ &= \frac{[\hat{V}(\hat{G}_{WT})](\hat{G}_{AT}) + [\hat{V}(\hat{G}_{AT})](\hat{G}_{WT})}{\hat{V}(\hat{G}_{AT}) + \hat{V}(\hat{G}_{WT})}. \end{aligned} \quad (10)$$

The variance of \hat{G}_{MT} could also be estimated by $\hat{V}(\hat{G}_{MT}) = W_A^2 \hat{V}(\hat{G}_{AT}) + W_W^2 \hat{V}(\hat{G}_{WT})$, where

$$W_A = \frac{\hat{V}(\hat{G}_{WT})}{\hat{V}(\hat{G}_{AT}) + \hat{V}(\hat{G}_{WT})} \text{ and } W_W = \frac{\hat{V}(\hat{G}_{AT})}{\hat{V}(\hat{G}_{AT}) + \hat{V}(\hat{G}_{WT})}. \quad (11)$$

In virtually all cases, \hat{G}_{MT} would lie between \hat{G}_{WT} and the arithmetic mean of \hat{G}_{WT} and \hat{G}_{AT} . This would occur because the among-twin-pair mean squares in all cases have one less degree of freedom and, more important, are almost invariably larger than the within-twin-pair mean squares. The \hat{G}_{MT} also has a theoretical advantage in that it is sensitive to the relative numbers of MZ and DZ twins, in distinction to other estimates of twin genetic variance.

In most instances, however, \hat{G}_{MT} would be very close to \hat{G}_{WT} because the among-twin-pair mean squares are usually several times greater than the within-pair-mean squares, and it would be more difficult to test the significance of \hat{G}_{MT} as compared to the F ratio (6) used to test significance of \hat{G}_{WT} . As a practical matter, therefore, it would appear to be of little more than theoretical value to

* At the suggestion of a reviewer, we verified empirically that if one modifies the matrix V of Haseman and Elston's weighted least-squares estimation procedure [1, pp. 15-16] by putting in the diagonals the approximate variances of the mean squares as given by Anderson and Bancroft [4, p. 319], then the estimate of their σ_g^2 obtained by applying their expression (26) (one iteration only) is exactly twice our \hat{G}_{MT} .

calculate \hat{G}_{MT} ; \hat{G}_{WT} would suffice as an estimate of genetic variance when the basic model holds.

If the F' test (3) gives evidence that $\sigma_{eMZ}^2 \neq \sigma_{eDZ}^2$, then the basic assumptions of the twin model are challenged and it must be modified to obtain an estimate of genetic variance. When $\sigma_{eMZ}^2 \neq \sigma_{eDZ}^2$ and we recalculate the expectations of \hat{G}_{WT} and \hat{G}_{AT} based upon the expected mean squares in table 2, we obtain the following equalities:

$$E(M_{AMZ} - M_{ADZ}) = [1/2 \sigma_a^2 + 3/4 \sigma_d^2 + (1-f) \sigma_i^2] + (\sigma_{eMZ}^2 - \sigma_{eDZ}^2), \quad (12)$$

$$E(M_{WDZ} - M_{WMZ}) = [1/2 \sigma_a^2 + 3/4 \sigma_d^2 + (1-f) \sigma_i^2] + (\sigma_{eDZ}^2 - \sigma_{eMZ}^2), \quad (13)$$

so that whenever $\sigma_{eMZ}^2 \neq \sigma_{eDZ}^2$, then \hat{G}_{AT} and \hat{G}_{WT} are biased by the difference between σ_{eMZ}^2 and σ_{eDZ}^2 .

However, as \hat{G}_{WT} and \hat{G}_{AT} are affected in opposite directions by inequality in σ_{eMZ}^2 and σ_{eDZ}^2 , an arithmetic mean of these two estimates is unbiased by a difference between σ_{eMZ}^2 and σ_{eDZ}^2 . We will call this estimate \hat{G}_{CT} because it is identical to Falconer's [5] among-twin-pairs component estimate of genetic variance. If σ_{AMZ}^2 and σ_{ADZ}^2 denote the variance components among twin pairs for monozygotic and dizygotic twins, respectively, it is well known that estimates of these components are given by $\hat{\sigma}_{AMZ}^2 = (M_{AMZ} - M_{WMZ})/2$ and $\hat{\sigma}_{ADZ}^2 = (M_{ADZ} - M_{WDZ})/2$.

Falconer [5, p. 184] proposes the difference $\hat{\sigma}_{AMZ}^2 - \hat{\sigma}_{ADZ}^2$ as one estimate of what we have called G_T . Substituting from above, it is readily verified that this is equivalent to $(\hat{G}_{AT} + \hat{G}_{WT})/2$, which is the proposed estimate \hat{G}_{CT} . Since it involves only the among-twin-pair variance components, \hat{G}_{CT} will not be affected by within-twin-pair components. The \hat{G}_{CT} is also equal to one-half the unweighted least-squares estimate of Haseman and Elston [1]. Hjortland [6] also pointed out that \hat{G}_{CT} is unbiased by differential environmental effects within the two twin types.

The hypothesis $E(\hat{G}_{CT}) = 0$ can be reduced to the hypothesis $E(M_{AMZ} + M_{WDZ}) = E(M_{ADZ} + M_{WMZ})$, which avoids minus signs, as Cochran [3] recommends. Thus, the significance of \hat{G}_{CT} can be tested by the ratio

$$F' = (M_{AMZ} + M_{WDZ}) / (M_{ADZ} + M_{WMZ}), \quad (14)$$

with approximate degrees of freedom calculated as in [3]. This should be a one-tailed F' test because if genetic variance is present the expected value of the numerator is greater than that of the denominator.*

* A Fortran program which calculates the various estimates of genetic variance and their variances and probability levels has been developed by one of us (K. W. K.) and is available upon request.

DISCUSSION

Six pieces of information are available to estimate genetic variance from the twin model, namely, the four mean squares in table 1 and the numbers of monozygotic and dizygotic twins. However, two assumptions, equations (4) and (5), are necessary, and neither can be tested using the twin model. The assumption that $C_{MZ} = C_{DZ}$ must hold because inequalities in these environmental covariance terms bias all estimates of genetic variance. For example, if C_{MZ} and C_{DZ} are positive and $C_{MZ} > C_{DZ}$, then M_{AMZ} and M_{WDZ} would be increased relative to M_{ADZ} and M_{WMZ} , respectively, inflating \hat{G}_{AT} and \hat{G}_{WT} by environmental effects. This appears to be the most serious potential flaw in the twin model, and we suspect that estimates of G_T are often thus inflated. The remaining assumption that $\sigma_{ge} = \sigma^*_{ge}$ is complex because genetic-environmental interaction may be logically partitioned into two covariance components [1]: σ_{ge} , the covariance between genetic and environmental effects within the same individual, and σ^*_{ge} , the covariance between the genotype of one twin and environmental effects on the other member of the twin pair. The σ^*_{ge} will only affect members of DZ twin pairs because there is no genetic variation between the two members of MZ sets. The genetic makeup of one twin could be expected to influence the environment of his cotwin either competitively or symbiotically. If there is a competitive relationship, one twin may be genetically equipped to successfully compete with the cotwin for environmental resources, causing less similarity in the twins and a negative σ^*_{ge} . In contrast, a positive σ^*_{ge} would be present when one twin is genetically influenced to seek different environments and by association the cotwin is exposed to these same new environments. A positive σ^*_{ge} may result in an environment beneficial to both (mutualism) or detrimental to both (synnecrosis). A positive σ^*_{ge} decreases variability within DZ twin sets and concomitantly increases variability among DZ twin sets, thus explaining the minus sign in $E(M_{WDZ})$ and the plus sign in $E(M_{ADZ})$ (table 1). In contrast, σ_{ge} contributes positively to the within- and among-DZ mean squares as well as the among-MZ mean square.

It is necessary, therefore, to assume that $\sigma_{ge} = \sigma^*_{ge}$ in order to estimate genetic variance free of the influence of these covariances. For example, if $\sigma_{ge} > \sigma^*_{ge}$, then the estimate of genetic variance will be biased upward. However, this does not seem to be a severe flaw in the model because σ_{ge} is evidence for genetic effects that may be modified by the environment (e.g., treatment). On the other hand, if $\sigma_{ge} < \sigma^*_{ge}$ it may become difficult to detect significant genetic variance. It is doubtful whether the twin model or any other population genetic model will detect or separate the components of genetic-environmental interaction before specific gene effects and environmental influences are identified for study.

Falconer's [5] within-twin-pair estimate of genetic variance and Haseman and Elston's alternative maximum-likelihood estimate of genetic variance [1, eq. (31)] are two examples of estimating genetic variance using only the within-twin-pair

mean squares (\hat{G}_{WT}). This approach at first seems wasteful because two of the four available mean squares and almost one-half of the available degrees of freedom are discarded. However, all of the estimates of genetic variance are exactly equal (\hat{G}_{AT} , \hat{G}_{WT} , and \hat{G}_{CT}) if the sums of the within- and among-mean squares of MZ and DZ twins are equal. In this instance \hat{G}_{WT} should suffice as an estimate because its significance is readily tested and it has a relatively small variance. When these sums of mean squares are significantly unequal by the F' test of equation (3), then the basic model should be modified. We have chosen to follow Kempthorne and Osborne [2] and postulate that the most likely causes of this situation are "competitive forces" that are different for monozygotic and dizygotic twins. They [2] discussed these competitive forces in detail and listed several causes including unequal partition of the cytoplasm in the case of MZ twins and intrauterine competition or competition between individuals after birth that could affect MZ or DZ twins. We have considered these competitive forces as environmental influences unique to each twin type and causing an inequality of environmental variance components ($\sigma_{eMZ}^2 \neq \sigma_{eDZ}^2$). This inequality would leave the arithmetic mean of \hat{G}_{AT} and \hat{G}_{WT} (\hat{G}_{CT}) as the only unbiased estimator of \hat{G}_T . Because inequalities in σ_{eMZ}^2 and σ_{eDZ}^2 may be relatively small compared to the sums of mean squares of MZ and DZ twins and yet seriously affect \hat{G}_{WT} or \hat{G}_{AT} , the most conservative approach would be to use \hat{G}_{CT} uniformly in twin studies. Unfortunately, \hat{G}_{CT} will almost always have a larger variance than \hat{G}_{WT} and may be overly conservative for routine use in detecting genetic variance. We would therefore suggest combining \hat{G}_{AT} and \hat{G}_{WT} into \hat{G}_{CT} only when there is evidence that the sums of mean squares of MZ and DZ twins are significantly different for the trait being studied ($P < .2$).

For traits where the sums of mean squares of MZ and DZ twins are unequal, further studies may lead to important information about sources of environmental variation. For example, the study of newborn twins may determine whether the discrepancy is due to pre- or postnatal factors.

The twin model will undoubtedly continue to be used to determine if there is evidence for significant genetic variance and to obtain an estimate of the magnitude of population genetic variance. However, if estimates of dominance (σ_d^2) and epistasis (σ_i^2) are not available from other family studies, then an estimate of population genetic variance may be obtained only by further assuming $\sigma_d^2 = 0$ and $\sigma_i^2 = 0$ and multiplying the estimates presented here by two.

SUMMARY

The general model and assumptions required to estimate genetic variance from monozygotic (MZ) and dizygotic (DZ) twins are reviewed. Using the among- and within-twin-pair mean squares, two independent estimates of genetic variance are obtained: \hat{G}_{WT} = within-DZ mean square — within-MZ mean square and \hat{G}_{AT} = among-MZ mean square — among-DZ mean square. This model holds only if the total mean squares of DZ and MZ twins differ only by chance fluctuation. For most cases, \hat{G}_{WT} is presented as an adequate measure of genetic variance,

but it may be refined by adjustment by a weighted mean with \hat{G}_{AT} . The weighted mean proposed was a minimum variance estimate in which \hat{G}_{AT} and \hat{G}_{WT} were combined according to the reciprocal of their variances.

If the total mean squares differ more than could be expected by chance, then environmental factors were postulated as being unequal for MZ and DZ twins and the arithmetic mean of \hat{G}_{WT} and \hat{G}_{AT} must be used as an unbiased estimate of twin genetic variance.

ACKNOWLEDGMENTS

The assistance of Dr. Manning Feinleib, Robert Garrison, and Richard King is gratefully acknowledged.

REFERENCES

1. HASEMAN JK, ELSTON RC: The estimation of genetic variance from twin data. *Behav Genet* 1:11-19, 1970
2. KEMPTHORNE O, OSBORNE RH: The interpretation of twin data. *Am J Hum Genet* 13:320-339, 1961
3. COCHRAN WG: Testing a linear relation among variances. *Biometrics* 7:17-32, 1951
4. ANDERSON RL, BANCROFT TA: *Statistical Theory in Research*. New York, McGraw-Hill, 1952
5. FALCONER DS: *Quantitative Genetics*. New York, Ronald, 1960
6. HJORTLAND MC: The effects of heredity and environment on nutrient intake of adult monozygotic and dizygotic twins. Doctoral diss. Univ. Minnesota, 1972